Hate speech classification in social media using emotional analysis

Ricardo Martins, Marco Gomes, José João Almeida, Paulo Novais, Pedro Henriques
Algoritmi Centre / Department of Informatics
University of Minho, Braga - Portugal
ricardo.martins@algoritmi.uminho.pt, {marcogomes, jj, pjon, prh}@di.uminho.pt

Keywords: Sentiment Analysis, Emotion Analysis, Natural Processing Language

Abstract: In this paper, we examine methods to classify hate speech in social media. We aim to establish lexical baselines for this task by applying classification methods using a dataset annotated for this purpose. As features, our system uses Natural Language Processing (NLP) techniques in order to expand the original dataset with emotional information and provide it for machine learning classification. We obtain results of 80.56% accuracy in hate speech identification, which represents an increase of almost 100% from the original analysis used as a reference.

1 Introduction

The proliferation of hate speech poses a new set of challenges. Notably, in the fast-paced and fragmented online discussion - in which many key-features are not present (such as gestures, facial expressions, intonation, etc.) the words used and writing styles can reveal information about our preferences, thoughts, emotions, and behaviours. Despite widespread recognition of the problems posed by such content, reliable solutions even for detecting hateful speech are lacking. In fact, hate speeches published and diffused via online environments have the potential to cause harm and suffering to individuals and lead to social disorder beyond cyberspace. Therefore, the detection of abusive language in user-generated online content has become an issue of increasing importance in recent years. Having automated techniques aim to programmatically classify text as hate speech, making its detection easier and, consequently, its mitigation.

One way to detect hate speech is using a lexicon-based approach, as presented by Gitari [4]. Meanwhile, as claimed by Davidson [2], lexical detection methods tend to have low precision because they classify all messages containing particular terms as hate speech. In our work, we investigate the problem of detecting hate speech online using lexical and emotional approaches.

2 Background

In this section, we present the definitions that are important to clarify in the problem of hate speech automatic classification. We analyse here different perspectives on the hate speech definition and also the work conducted so far in, mainly, lexicon-based hate speech automatic detection/classification.

2.1 Introduction to Hate Speech

"Hate speech" is an emotive concept, and there is no universally accepted definition of it in international human rights law. Many would claim they can identify "hate speech" where they see it, but the criteria for doing so are often elusive or contradictory. Therefore, for Brown [1] the idea that the concept "hate speech" might be a complex concept, composed of two basic concepts *hate* and *speech*. According to the author, it can be split into two main components:

- Hate: the intense and irrational emotion of opprobrium, enmity and detestation towards an individual or group, targeted because of their having certain - actual or perceived – protected characteristics (recognised under international law). "Hate" is more than mere bias, and must be discriminatory. Hate is an indication of an emotional state or opinion, and therefore distinct from any manifested action.
- Speech: any expression imparting opinions or ideas – bringing a subjective opinion or idea to an external audience. It can take many forms: written, non-verbal, visual or artistic, and can be disseminated

through any media, including internet, print, radio, or television.

Beyond these two basic elements, and to put simply, "hate speech" is any emotional expression of hate towards people. Based on this, we define, as a working definition, the hate speech in the scope of this work as:

> Any emotional expression imparting opinions or ideas – bringing a subjective opinion or idea to an external audience- with discriminatory purposes. It can take many forms: written, non-verbal, visual, artistic, and may be disseminated through any media, including internet, print, radio, or television.

2.2 Related Work

In recent years, at the same time as the hate speech concept has become more popular, several works have been published related to the identification, detection, and characterisation of hate speech and its actors. However, few datasets have been collected, annotated, and released by other researchers around abusive behaviour on social media. Indeed, there is a general lack of systematic monitoring, documentation and data collection of hate content. It is a fundamental problem that limits much of the results of the latest studies.

Despite this limitation, studies on computational methods to hate speech detection have been growing mainly focusing on adapting strategies in text mining to the specific problem of automatic hate speech detection. When examining the methods to detect hate speech in social media, the most common approach found consists in building Machine Learning models for hate speech classification. The majority of the studies try to adapt strategies already known in text mining to the specific problem of automatic hate speech detection, e.g. using dictionaries and lexicons. There are some works which served as inspiration for this approach. For example, one exciting work aimed to establish lexical baselines for classification by applying supervised classification methods using a dataset annotated for this purpose is presented by Malmasi and Zampieri [7]. They achieve to obtain a 78% accuracy in identifying posts across three classes (Hate, Offensive and Ok) using an approach based on N-Gram and linear SVM to perform multi-class classification. It contributed to the idea of preprocessing a text prior to the classification using machine learning. Davidson [2], using a crowd-sourced hate speech lexicon and a trained multi-class classifier to distinguish between hate speech from another offensive language (when differentiation is more difficult) presented a work which concluded that racist and homophobic tweets are more likely to be classified as hate speech, but that sexist tweets are classified as offensive. This work contributed as the source of the dataset used in our tests.

Sentiment analysis is a valuable tool that helps to increase machine learning classification, as demonstrated by Martins [9], which used an approach based in lexical analysis and machine learning classification to increase the authorship identification, contributing with the idea of using emotional labels as dimensions in the classifier, using the same approach presented by Xu [14] who apply sentiment analysis to detect bullying in tweets, still scarce.

3 Text Processing & Data analysis

Social media is a rich source of information about opinions. According to Foux [3], "social media is perceived by consumers as a more trustworthy source of information regarding products and services than corporate-sponsored communications transmitted via the traditional elements of the promotion mix." Due to the enormous amount of information and opinions available in social media, unfortunately, it is difficult to filter what users post individually, which facilitates the spread of hate speech in these media since often the reaction to the posting is delayed.

On the other hand, it is undeniable that the hate speech is loaded with emotions. Since these emotions in the text may represent an emotional pattern characteristic of this type of discourse, an alternative to help identify this type of content would be to analyse the emotional profile of the comments.

3.1 Sources validation

To detect hate speech, it is mandatory first know what hate speech is and which features are relevant in hate speech. For this purpose, it was analysed the hate speech dataset provided by Davidson and Warmsley [2], which provides a set of 24782 tweets classified in hate speech (1430), offensive language (19190) or neither (4162). These tweets were manually classified, so the differentiation between each classification was subjective, according to the reviewer interpretation.

The dataset was deconstructed using Natural Processing Language (NLP) techniques to identify which words into the set of all texts classified as hate speech were more relevant and more frequent and whether that content pointed to hate speech.

After all the sentences had their stopwords removed, it was applied TF-IDF to rank the most relevant words in

the dataset and count all words to rank the most frequent words. Visually it is easy to figure out that the most well-ranked words contain a rather marked hatred, as can be seen in Fig. 1 where the most well-classified words are more significant and located in the middle of the word cloud while the worst classified words are smaller and located in the edges, however, it does not guarantee that they are hate speech. So, a way to ensure that such words are hate speech is to consult their existence in a specific lexicon. For this action, it was used the lexicon provided by Hatebase[1] containing a set of words considered hate speech.

The objective of this analysis is to assure that the dataset contains hate speech. To achieve this objective, two different analysis were performed: a TF-IDF to detect the most relevant words in the texts and counting of a bag-of-words to identify the most frequent words. Thereby, according to Table 1, the more significant is the relevant words set, the smaller is the number of these words in the hate speech lexicon, indicating that the most relevant words in the dataset tend to be words of hatred. Also, regardless of the number of frequent words, the percentage of their existence in the hate lexicon is less than 50%, indicating that their use as a metric is not as reliable as the most relevant words.

Table 1: Dataset's words existence in hate lexicon

	Relevant words	**Frequent words**
Top 10	50%	40%
Top 25	60%	44%
Top 50	42%	34%
Top 100	31%	26%
Top 200	22%	18%
Top 500	12%	10%

So, according to this analysis, the dataset can be considered as a good source for training the model to detect the hate speech, because in their sentences classified as hate speech, there more than 50% of relevant words from hate speech classified sentences identified as hate speech in Hatebase's lexicon.

3.2 Emotion models for detecting emotion in text

In the literature, there are different theories to model emotions, their associated behaviours and discuss how emotions are elicited in our cognitive system.

According to Scherer [13], the current models of emotion can be divided into four major groups: dimensional models, discrete emotions models, meaning-oriented models and componential models. For the authors, the primary focus of dimensional models is subjective feeling, discrete emotion models focus in motor expression or adaptive behaviour patterns, meaning models are aimed to a verbal description of subjective feelings, and componential models focus in a link between emotion-antecedent evaluation and differentiated reaction patterns.

While the objective of this work is identifying the emotions contained in hate speech and how they can contribute to its identification, in this study will be only considered the discrete emotions model.

The **discrete emotional models** consider that every emotion is composed of universally displayed and recognized basic emotions, as *happiness*, *anger*, *sadness*, *surprise*, *disgust*, and *fear*, for instance. One of the main advantages of discrete models is that, through psychophysical experiments, the perception of emotions by human beings is discrete.

Discrete models group emotions into categories and assume that they are independent. In the literature, among the discrete models, a well-known model is the so-called wheel of basic emotions, proposed by Plutchik [11]. This model proposes the existence of eight basic emotions: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*. The secondary emotions around the perimeter arise from the blending of two primary emotion which are close to each other. The emotional intensity is represented by the colour intensity, where the most intense levels are solid and vice versa.

To define the emotional model, all phrases classified as hate speech in the dataset were analysed through a lexicon-based approach, consisting in comparing the labelled emotion contained into the EmoLex [10] lexicon against the words existent in the phrase.

Considering the texts available into this dataset, it was possible to define the hate speech emotional model, as presented in Fig. 2, containing the values of each basic emotion.

Moreover, in this analysis, the positive and negative polarities of the texts have reached an average of 0.32 and 0.68 respectively. It means that for every positive word in a hate speech, in general, there are two negative words.

Hence, when applying the Person's correlation coefficient (r^2) between polarities and basic emotions, as presented in Table 2, it is possible to figure out which basic emotion is related to positive and negative polarities.

Another analysis made was concerned in the distribution of hate speech. For this objective, when applying the r^2 to the emotional values for each hate speech text

[1] http://www.hatebase.org

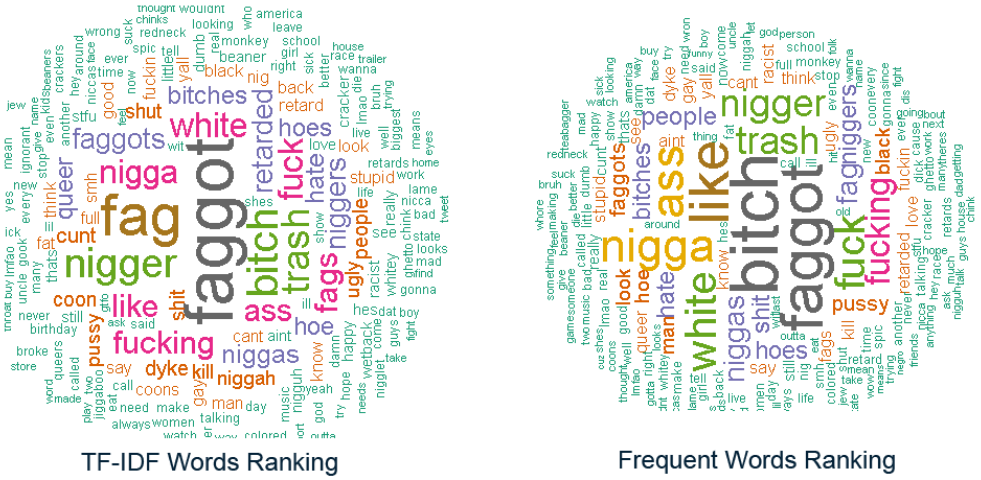

Figure 1: Wordcloud of hate speech words

Table 2: Correlation between polarities and emotions

Polarity	Avg. Value	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Negative	1.08	0.63	0.17	0.70	0.60	0.11	0.68	-0.11	0.17
Positive	0.51	0.04	0.57	-0.15	0.11	0.75	0.14	0.30	0.69

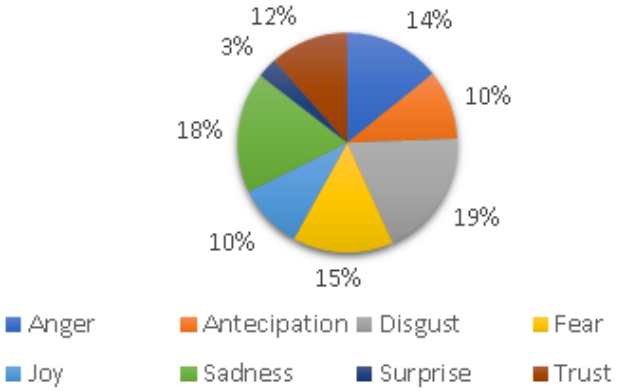

Figure 2: Hate speech emotional model

against the emotional hate speech model it was possible to identify that in the labelled hate speech:

- Only 2.67% (26 of 975) of the messages have a very strong correlation with the model;
- 4.62% (45 of 975) of the messages have a strong correlation with the model;
- 6.87% (67 of 975) of the messages have a moderated correlation with the model;
- 6.46% (63 of 975) of the messages have a weak correlation with the model;
- 4.41% (43 of 975) of the messages have a non-linear relationship with the model;
- 43.49% (424 of 975) of the messages have a negative correlation with the model;
- 31.49% (307 of 975) of the messages do not have emotions detected.

These analysis points some interesting informations:

1. Based on values fewer than 0.5 for both polarities - indicating a weak correlation - *surprise* can be interpreted as a neutral emotion in hate speech;
2. All sentences can be grouped into two groups: positive {*anticipation, joy, trust*} and negative {*anger, disgust, fear, sadness*}.
3. Once the negative words occur in 2/3 of hate speech texts, the most critical emotions to identify hate speech are *anger, disgust, fear* and *sadness*.
4. It is necessary to expand the emotional lexicon to consider new words. Since that almost 1/3 of sentences do not have their emotions identified when expanding the lexicon will increase the number of emotional words identified, and consequently, the correlation values will increase.

4 Machine learning analysis

Once identified the average of each emotion in a hate speech, the next step was to train a model to classify the hate speech. For this purpose, a new dataset was created based on Davidson's dataset.

4.1 Dataset creation

To perform this step, it was created a new dataset, containing 975 preprocessed tweets previously categorised in Davidson's dataset for each category (hate speech, offensive speech and neither). Furthermore, it was added a flag indicating if the sentence contains words identified in the Hatebase lexicon as hate speech, and using the NRC Intensity lexicon as source, all sentences are analysed and the intensity of the emotion *anger* is calculated. Later, through a Natural Language Processing (NLP) pipeline, the number of words in the text was decreased, remaining only the ones that bring relevant information.

The pipeline used for text decreasing contains five steps: n-Gram identification, tokenisation, stopwords removal, part of speech tagging and named entity removal, as presented in Fig. 3.

Using the Stanford Core NLP toolkit [8] for these tasks, the preprocessing is divided into three parallel tasks. This is important because both part of speech tagging and named entity recognition need the text in the original format to identify the information.

The preprocessing begins with the n-Gram identification, where a predefined set of n-Grams are identified in the text and labelled to be interpreted as a single word. Moreover, the most frequency bi-Grams and tri-Grams (pairs of words and triples of words, respectively) are identified to be manually evaluated and added in a stopwords list.

In the next step, the tokenizer splits the text in a list of words (tokens). So, the pipeline begins three parallel processes:

- The tokens are syntactically analysed in part of speech, where the nouns, verbs, adverbs and adjectives are identified and stored for future purposes;
- The tokens contained in a predefined (and updated with N-Grams information) stopwords list are removed;
- The tokens in named entity process are analysed to identify names (persons, locations or organisations) and discard them.

Finally, the resulting tokens in common for these processes are stored, and the emotions from each preprocessed text are identified through a process in R [12] which queries the EmoLex lexicon and identifies the basic emotions according to the Plutchik's model.

Using the Syuzhet package for R [6] and the NRC-Intensity lexicon the intensity of the emotion *anger* was calculated. A surprising confirmation in this analysis is related to how to differentiate the hate speech and offensive language. In our analysis, the mean of the anger intensity have shown that hate speech is less intense than offensive language, having a mean of 0,29 against 0,51 from offensive language and 0,124 for neither.

So, the final dataset is composed of 14 dimensions: the classification, the original text message, eight basic emotion from Plutchik's model, the polarities (positive and negative) and a flag indicating the existence of a word in tweet contained into the Hatebase lexicon, and the intensity of emotion *anger*, as shown in Table 3.

4.2 Feature selection

In machine learning, feature selection is the use of specific variables (dimensions) or data points to maximise the efficiency of an algorithm. To determine which dimensions are more useful to classify the tweets, the feature selection was applied to the new dataset.

Using Weka [5], the InfoGainAttributeEval as attribute evaluator, the Ranker search method and the full dataset as the training set, the most relevant attributes to select the class (hate speech, offensive speech or neither) in the dataset are identified as presented in Table 4.

These results reinforce the analysis presented in subsection 3.1, showing that all negative emotions are well-ranked in text classification than the positives, and their importance in helping to identify hate speech in texts.

4.3 Data classification

For data classification measurements, it was considered as comparison basis the precision and recall results presented by Davidson [2], whose values are 0.41 and 0.61 respectively.

In our tests, the information used to evaluate the approach was from the new dataset generated previously, containing the classification, text, indication of words in the Hatebase lexicon and emotional labels (emotions and intensity).

For the text classification, it was used the Weka [5] with the most relevant algorithms for text classification, such as SVM, Naive Bayes and Random Forest and a ten folds cross-validation for results validation, in default parameters configuration for each algorithm. We choose the default configuration approach to avoid bias for better tunning among the algorithms.

In all cases, the precision and recall results on predicting hate speech were superior to Davidson results, with Random Forest algorithm having the best precision, as presented in Table 5, where PR and RC indicate "Precision" and "Recall" respectively. The overall

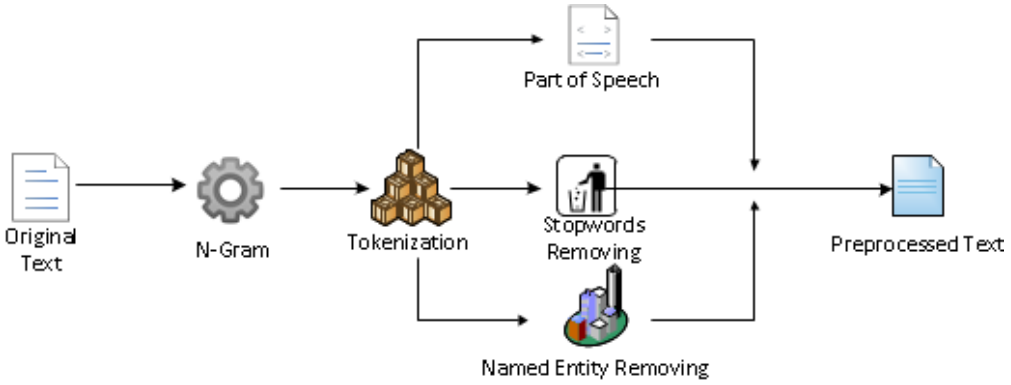

Figure 3: Preprocessing pipeline

Table 3: Training dataset

Class	Tweet	FreeOfHateWord	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust	Negative	Positive	Intensity
Offensive	RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. \& as a man you should always take the trash out...	1	1	0	1	0	0	2	0	0	2	1	0.64
Neither	RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!	1	1	0	2	1	0	1	0	0	3	0	0.12
Neither	RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit	1	2	1	2	1	0	2	0	0	3	0	0.095
Neither	RT @C_G_Anderson: @viva_based she look like a tranny	1	0	0	0	0	0	0	0	0	0	0	0.089
Neither	RT @ShenikaRoberts s: The shit you hear about me might be true or it might be faker than the bitch who told it to ya	1	2	0	2	1	0	1	0	0	2	0	0.124

Table 4: Dimension's importance on text classification

Rank	Dimension
0.11765	Intensity
0.05979	Anger
0.04294	Fear
0.05241	FreeOfHateWord
0.0504	Negative
0.03478	Disgust
0.01541	Sadness
0.01069	Positive
0.00598	Trust
0.00449	Surprise
0	Anticipation
0	Joy

results for each algorithm, despite Davidson did not provide this information for comparison, can be considered as successful, once that the worst result was obtained for Naive Bayes algorithm, which obtained an average of 71.33%, and the best result was obtained for SVM, with an average of 80.56%.

5 Conclusion

In this paper we presented a combination of lexicon-based and machine learning approaches to predict hate speech contained in a text, using an emotional approach through sentiment analysis.

Using the emotional information contained in text helps to increase the accuracy on hate speech detection. This claiming is based on the successful precision rate grown from 41% in the original research to 80.64% in our tests. This improvement of almost 100% can be interpreted as a successful result as our proposal.

Nevertheless, our analysis still has limitations that lead to exciting future research directions. Firstly, it is reasonable to question the definition of hateful content, in the sense that it is not clear what is the threshold a published text shared in social media has to violate to be considered hateful due to the subjectivity of the definition of hate-speech. Secondly, this work does not address the issue of users characterisation and their potential use of code to overcome anti-hate speech policies and automatic detection systems. Thirdly, since the words used in hate speech change rapidly - new words creation, expressions used locally or within a given context - it is a somewhat arduous task to be up to date with the new expressions used.

As future work, it is planned the creation of a classification module for new emotional words, to increase the ability to analyse new words without the dependence of specialised and updated lexicons and consequently increase the prediction of hate speech. Another line of future work is to explore computational strategies and

Table 5: Detailed algorithms results

	Naive Bayes		SVM		Random Forest	
Class	**PR**	**RC**	**PR**	**RC**	**PR**	**RC**
Hate Speech	0.701	0.525	0.768	0.736	0.816	0.646
Offensive Language	0.724	0.785	0.824	0.77	0.781	0.767
Neither	0.714	0.834	0.825	0.913	0.756	0.926

approaches to characterise and monitor user-centric content in social media.

Acknowledgements

This work has been supported by COMPETE: POCI-01-0145-FEDER-0070 43 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/ 00319/2013.

REFERENCES

[1] Alexander Brown. What is hate speech? part 1: The myth of hate. *Law and Philosophy*, 2017.

[2] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[3] Graeme Foux. Consumer-generated media: Get your customers involved. *Brand Strategy*, 8(202):38–39, 2006.

[4] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[5] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[6] Matthew L. Jockers. *Syuzhet: Extract Sentiment and Plot Arcs from Text*, 2015.

[7] Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*, 2017.

[8] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[9] Ricardo Martins, José Almeida, Pedro Henriques, and Paulo Novais. Increasing authorship identification through emotional analysis. In *Advances in Intelligent Systems and Computing*, volume 745, pages 763–772. Springer International Publishing, 2018.

[10] Saif Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[11] Robert Plutchik. *Emotion: A psychoevolutionary synthesis*. Harpercollins College Division, 1980.

[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[13] Klaus R Scherer. Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162, 2000.

[14] Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, 2012.